UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MASSACHUSETTS

SINGULAR COMPUTING LLC, Plaintiff, v. GOOGLE LLC, Defendant.	Civil Action No. 1:19-cv-12551 FDS Hon. F. Dennis Saylor IV

STATEMENT OF UNDISPUTED MATERIAL FACTS IN SUPPORT OF DEFENDANT GOOGLE LLC'S MOTION FOR SUMMARY JUDGMENT OF NON-INFRINGEMENT

Pursuant to Fed. R. Civ. P. 56 and Local Rule 56.1, Defendant Google LLC ("Google"), submits this Statement of Undisputed Material Facts in Support of its Motion for Summary Judgment of Non-Infringement.

The following paragraphs set forth facts[1] for which there is no genuine issue to be tried, as well as citations to the record establishing such facts. For the Court's convenience, relevant portions of the record that provide support for such facts are quoted in footnotes below:

I. BACKGROUND

1. Plaintiff Singular Computing LLC ("Singular") asserts two claims ("Asserted Claims") against Google: dependent claim 53 of U.S. Patent No 8,407,273 (the "'273 patent") and dependent claim 7 of U.S. Patent No. 9,218,156 (the "'156 patent"). *See* Dkt. 410-3 (copy of Singular's August 11, 2022 supplemental infringement claim chart for the '156 patent, filed with Google's motion to strike expert report of Sunil Khatri); Dkt. 410-4 (copy of Singular's August 11, 2022 supplemental infringement chart for the '273 patent, also filed with Google's motion to strike).

2. Singular's technical expert on infringement, Dr. Sunil Khatri, contends that two versions of Google's Tensor Processing Unit ("TPU") boards infringe the Asserted Claims: version 2 of the TPU board ("TPUv2") and version 3 of the TPU board ("TPUv3"). Ex. 1[2] (Khatri Report) ¶¶ 91–92.[3]

---

[1] Google's identification of undisputed facts herein is made only for purposes of the present Motion for Summary Judgment of Non-Infringement. Google reserves its right to contend that some or all of the facts herein are not adequately established by Defendant Singular Computing LLC.

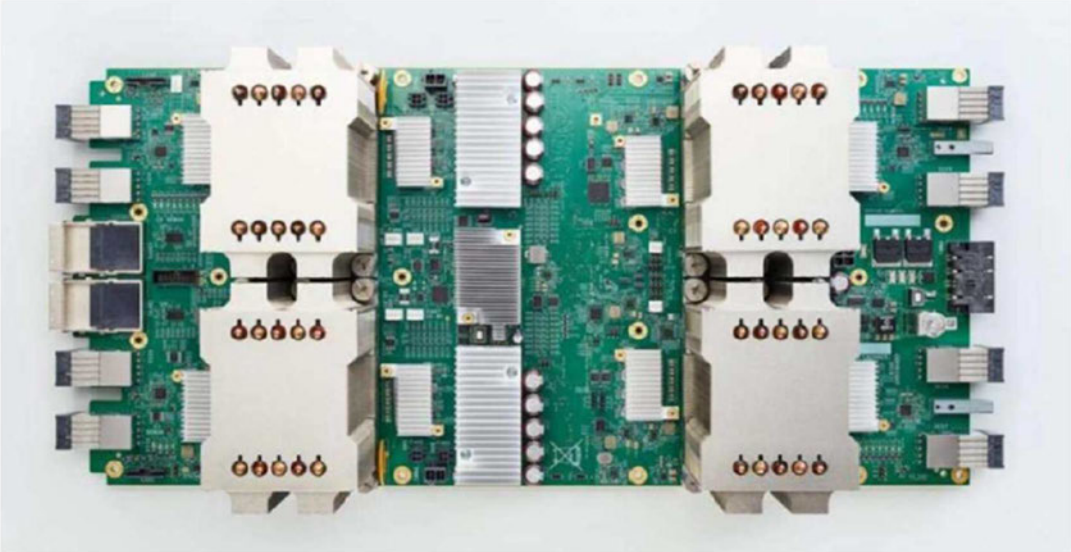
[2] Exhibits cited herein are attached to the Declaration of Vishesh Narayen in Support of Defendant Google LLC's Motion for Summary Judgment of Non-Infringement, filed herewith.

[3] "The accused TPUv2 product, known internally as 'Jellyfish,' comprises a circuit board to

II. THE ACCUSED TENSOR PROCESSING UNIT BOARDS

3. The accused TPU boards are application-specific integrated circuit boards designed to accelerate machine-learning tasks. Ex. 1 ¶ 71.[4]

4. As an example, the following image shows a TPUv2 board. Ex. 1 ¶ 82.



A. TPUv2 Board

5. Each TPUv2 board comprises four integrated circuits known as "Jellyfish Chips," or "JFCs," that are attached to the board. Ex. 1 ¶ 91.[5]

6. Each Jellyfish Chip (JFC) contains two "Tensor Cores," for a total of eight Tensor Cores per TPUv2 board. Ex. 1 ¶ 80.[6]

---

which four chips ('Jellyfish Chips' or 'JFCs') are attached." Ex. 1 ¶ 91.

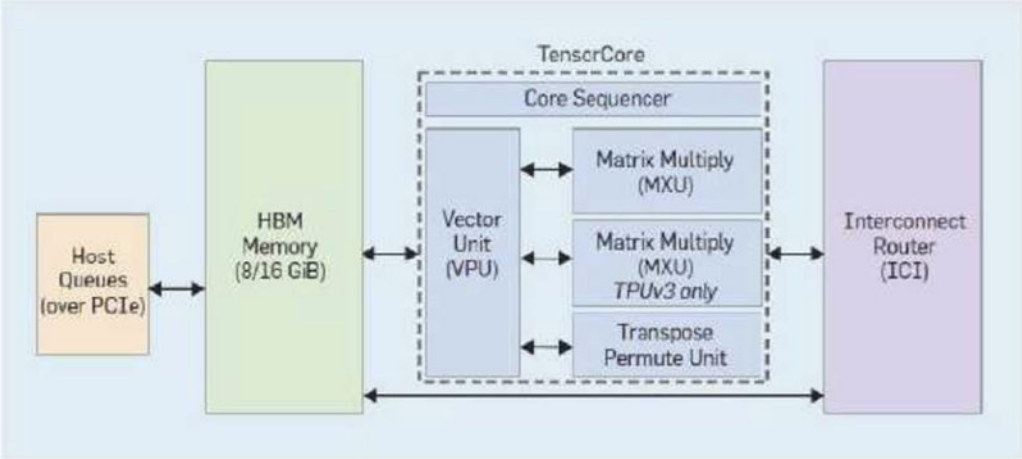
"The accused TPUv3 product, known internally as 'Dragonfish,' comprises a circuit board to which four chips ('Dragonfish Chips' or 'DFCs') are attached." Ex. 1 ¶ 92.

[4] "The TPUs are application-specific circuit boards, containing ASIC accelerators optimized specifically for machine learning operations." Ex. 1 ¶ 71.

[5] "The accused TPUv2 product, known internally as 'Jellyfish,' comprises a circuit board to which four chips ('Jellyfish Chips' or 'JFCs') are attached." Ex. 1 ¶ 91.

[6] "The TensorCores of the TPUv2 are included within chips ('Jellyfish Chips' or 'JFCs') that are attached to a circuit board. Each JFC includes two TensorCores." Ex. 1 ¶ 80.

7. The following block diagram illustrates one Tensor Core and some of its hardware components (in the dashed-line box). Ex. 1 ¶ 78.



8. Each Tensor Core in a JFC has one Vector Processing Unit ("VPU"), one Matrix Multiply Unit ("MXU"), and a Core Sequencer, for a total of eight MXUs, eight VPUs, and eight Core Sequencers on each TPUv2 board. Ex. 1 ¶¶ 94–95, 226.[7]

9. The MXU is capable of performing large matrix multiplication on 16-bit floating point values in the "bfloat16" (BF16) format. *See* Ex. 1 ¶ 102.[8]

10. Each VPU on a JFC contains 256 float conversion circuits ("rounding circuits"), each of which converts a 32-bit floating point value (FP32) to a 16-bit floating point value (BF16). Ex. 2 (Walker Report) ¶ 217;[9] Ex. 3 (Phelps Dep. Tr.) at 55:22–56:9, 92:22–94:3, 94:6–

---

[7] "Each Tensor Core contains a VPU, a Core Sequencer, and at least one MXU." Ex. 1 ¶ 94. "Specifically, Tensor Cores in the TPUv2 each contain a single MXU . . . ." Ex. 1 ¶ 95. "Each TPUv2 device has 1 MXU per Tensor Core, 2 Tensor Cores per JFC chip, and 4 JFC chips per TPUv2 board . . . for a total 8 MXUs . . . ." Ex. 1 ¶ 226.

[8] "Specifically, the 'MXUs' (or Matrix Multiply Units) within the Accused Products 'typically perform[] multiplications at the reduced precision of bfloat16.'" Ex. 1 ¶ 102.

[9] "Rather, each VPU Float Conversion circuit, of which there are 256 in JFC for its single MXU . . . ." Ex. 2 ¶ 217.

22.[10] Because the TPUv2 board has eight VPUs, each TPUv2 board comprises 2,048 rounding circuits. *Id.*

11. Dr. Khatri refers to the 256 rounding circuits as "precision-reducer circuits." Ex. 1 ¶ 231.[11]

12. Dr. Khatri states in his report that the 256 rounding circuits are in the MXU rather than the VPU, citing deposition testimony of Andrew Phelps, but Mr. Phelps stated that these circuits are "[i]n the VPU." Ex. 1 ¶ 231; Ex. 3 at 94:6–9.[12] Dr. Khatri admitted that "it's possible" the rounding circuits are "in the VPU." Ex. 4 (Khatri 3/23/23 Dep. Tr.) at 219:7-18.[13]

---

[10] [redacted] Ex. 3 at 55:22–56:9. [redacted] Ex. 3 at 92:22–94:3. [redacted] Ex. 3 at 94:6–22.

[11] "There are 256 precision-reducer circuits in the MXU." Ex. 1 ¶ 231.

[12] "There are 256 precision-reducer circuits in the MXU. *See* Phelps 94:16-22." Ex. 1 ¶ 231. "Q: Now in the diagram that you drew at the top of Exhibit 8, where are these rounding circuits physically located? A: In the VPU." Ex. 3 at 94:6–9.

[13] "Q: Okay. And those [rounding] circuits are physically located in the vector processing unit; is that right? . . . . A: You know, let's see. My memory of this might be failing me. I thought they were in the MXU, but they possibly could be in the – sorry, and the MXU is the multiplication unit. But it's possible that they [the rounding circuits] were in the VPU, or the vector processing unit." Ex. 4 at 219:7–18.

B. TPUv3 Board

13. The TPUv3 board is structurally similar to the TPUv2 board. Ex. 1 ¶ 83.[14]

14. Each TPUv3 board comprises four integrated circuits known as "Dragonfish Chips," or "DFCs," that are attached to the board. Ex. 1 ¶ 92.[15]

15. Each Dragonfish Chip (DFC) contains two Tensor Cores, for a total of eight Tensor Cores per TPUv3 board. Ex. 1 ¶ 93.[16]

16. Each of the eight Tensor Cores in the TPUv3 board contains two MXUs, for a total of sixteen MXUs on each TPUv3 board. Ex. 1 ¶ 227.[17]

17. Each VPU in the TPUv3 board has 512 rounding circuits. Ex. 2 ¶ 217;[18] Ex. 3 at 94:16–18.[19] Because the TPUv3 board has eight VPUs, each TPUv3 board comprises 4,096 rounding circuits. *Id.*

III. SINGULAR'S INFRINGEMENT THEORY

18. Singular alleges that Google infringes claim 53 of the '273 patent and claim 7 of the '156 patent.

---

[14] "Google designed and implemented a third-generation TPU (the TPUv3 or 'Dragonfish'), that includes many of the same components as the TPUv2 arranged in fundamentally the same way." Ex. 1 ¶ 83.

[15] "The accused TPUv3 product, known internally as 'Dragonfish,' comprises a circuit board to which four chips ('Dragonfish Chips' or 'DFCs') are attached." Ex. 1 ¶ 92.

[16] "Each JFC contains two 'Tensor Cores;' each DFC also contains two Tensor Cores." Ex. 1 ¶ 93.

[17] "Each TPUv3 device has 2 MXUs per Tensor Core, 2 Tensor Cores per DFC chip, and 4 DFC chips per TPUv3 board . . . for a total of 16 MXUs . . . ." Ex. 1 ¶ 227.

[18] "Rather, each VPU Float Conversion circuit, of which there are . . . 512 in DFC . . . ." Ex. 2 ¶ 217.

[19] "Q: So in a Dragonfish chip, there would be 512 rounding circuits, correct? A: To the best of my recollection." Ex. 3 at 94:16–18.

19. Asserted claim 53 of the '273 patent depends from claim 43, which in turn depends from independent claim 36. Ex. 5 ('273 patent) at col. 31–32. Incorporating the limitations of the earlier claims from which it depends, asserted claim 53 reads:

> **36.** A device:
>
> comprising at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value,
>
> wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/65,000 through 65,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X % of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input;
>
> wherein the number of LPHDR execution units in the device exceeds the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.
>
> **43.** The device of claim 36, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.
>
> **53.** The device of claim 43, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000.

Id.

20. Asserted claim 7 of the '156 patent depends from claim 3, which in turn depends from claim 2, which in turn depends from independent claim 1. Ex. 6 ('156 patent) at col. 29–30. Incorporating the limitations of the earlier claims from which it depends, asserted claim 7 reads:

> **1.** A device comprising:

at least one first low precision high dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value,

wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/65,000 through 65,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; and

at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit.

2. The device of claim 1, wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine.

3. The device of claim 2, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.

7. The device of claim 3, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000.

Id.

21. Dr. Khatri reads the "device" recited in the Asserted Claims on the TPUv2 and TPUv3 boards. Ex. 1 ¶¶ 90–92.[20]

[20] "First, the accused TPUv2 and TPUv3 products are both 'devices,' as the preamble recites." Ex. 1 ¶ 90.

"The accused TPUv2 product, known internally as 'Jellyfish,' comprises a circuit board to which four chips ('Jellyfish Chips' or 'JFCs') are attached." Ex. 1 ¶ 91.

"The accused TPUv3 product, known internally as 'Dragonfish,' comprises a circuit board to which four chips ('Dragonfish Chips' or 'DFCs') are attached." Ex. 1 ¶ 92.

22. The Asserted Claims require "at least one . . . low-precision high dynamic range (LPHDR) execution unit" (LPHDR EU). Ex. 5 at cl. 36; Ex. 6 at cl. 1.

23. The Asserted Claims also require that "the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide." Ex. 5 at cl. 43; Ex. 6 at cl. 3.

24. According to Dr. Khatri, each claimed LPHDR EU in the accused TPUv2 and TPUv3 boards comprises: (i) two rounding circuits; and (ii) a BF16 multiplication circuit in the MXU. Ex. 1 ¶¶ 140, 230, 289.[21] Dr. Khatri created the following diagram to illustrate the

[21] "Specifically, the LPHDR EU comprises the precision-reducer circuits that convert each of the FP32 input signals into low-precision BF16 signals . . . and the BF16 multiplication circuit . . . ." Ex. 1 ¶ 140.

"[T]he two precision-reducer circuits (the first stage of the pipeline) of the LPHDR execution unit are converting two new FP32 input signals into BF16 signals for the next LPHDR multiplication operation." Ex. 1 ¶ 230.

"[T]he components of the TPUv2 and TPUv3 that constituted the claimed 'LPHDR execution units,' . . . include, among others, the precision-reducer circuits that convert FP32 input signals to BF16 signals, and BF16 multipliers that multiply BF16 signals to produce a FP32 product." Ex. 1 ¶ 289.

components of what he asserts are the LPHDR EUs in the accused TPUv2 and TPUv3 boards. Ex. 1 ¶¶ 140–141.[22]

25. Dr. Khatri opines that each LPHDR EU in the accused TPU boards is "pipelined" with "two pipeline stages." Ex. 1 ¶ 230.[23] According to Dr. Khatri, the first pipeline stage consists of "the two precision-reducer circuits" (two rounding circuits) that convert two FP32 input signals into BF16 signals; the second pipeline stage comprises the BF16 multiplication circuit that multiplies BF16 signals and produces an output signal for the "next LPHDR multiplication operation." *Id.*

---

[22] "The above figure, which I created by combining Mr. Phelps's sketch and GOOG-SING-00236144 at 51, shows the components of the LPHDR EU of the accused TPU devices. They include the precision-reducer circuits (the boxes labeled 'R') [redacted] and the BF16 multiplication circuit (shown above as a circle marked with '✕')." Ex. 1 ¶ 141.

[23] "The LPHDR execution units of the accused TPUv2 and TPUv3 devices are pipelined in this fashion, using two pipeline stages. While the BF16 multiplier (the second stage of the pipeline) of an LPHDR execution unit is multiplying BF16 signals and producing an output signal for the current LPHDR multiplication operation, the two precision-reducer circuits (the first stage of the pipeline) of the LPHDR execution unit are converting two new FP32 input signals into BF16 signals for the next LPHDR multiplication operation." Ex. 1 ¶ 230.

26. The Asserted Claims require that the number of LPHDR EUs in an accused device "exceeds by at least one hundred" the number of "execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide." Ex. 5 at cl. 43; Ex. 6 at cl. 3.

27. Dr. Khatri opines that the number of "execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide" in each accused TPUv2 and TPUv3 board is 8,200. Ex. 1 ¶¶ 234–235.[24]

28. Specifically, Dr. Khatri opines that each Tensor Core contains 1,025 "execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide" because [redacted] Ex. 1 ¶ 234.[25]

29. Under Singular's infringement theory, there must be at least 8,300 LPHDR EUs in each TPUv2 and TPUv3 board in order to satisfy the requirement of the Asserted Claims that "the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide."

30. There are only 2,048 rounding circuits on each TPUv2 board. *See* ¶ 10, *supra*.

---

[24] "There are only 1,025 execution units in each Tensor Core that are adapted to perform 32-bit floating point multiplication [redacted] Ex. 1 ¶ 234.
"This makes 8,200 execution units in each TPUv2 and TPUv3 board (1025 × 8) capable of 32-bit multiplication . . . ." Ex. 1 ¶ 235.

[25] "There are only 1,025 execution units in each Tensor Core that are adapted to perform 32-bit floating point multiplication [redacted] Ex. 1 ¶ 234.

31. There are only 4,096 rounding circuits on each TPUv3 board. *See* ¶ 17, *supra*.

32. Dr. Khatri asserts that each TPUv2 board contains 131,072 LPHDR EUs because a TPUv2 board performs that many "LPHDR multiplication operations" per clock cycle. Ex. 1 ¶¶ 224–226, 228.[26]

33. Dr. Khatri asserts that each TPUv3 board contains 262,144 LPHDR EUs because a TPUv3 board performs that many "LPHDR multiplication operations" per clock cycle. Ex. 1 ¶¶ 224–225, 227–228.[27]

34. A clock cycle is a measure of time which reflects the pace at which units within a computer operate. Ex. 1 ¶ 62 n.2.[28]

---

[26] "Every clock cycle, MXUs of the accused TPUv2/3 devices complete 16,384 (128 x 128) LPHDR multiplication operations, with 16,384 separate output signals produced by 16,384 BF16 multiplier circuits." Ex. 1 ¶ 224.

"Thus, there 16,384 LPHDR execution units per MXU." Ex. 1 ¶ 225.

"Each TPUv2 device has 1 MXU per Tensor Core, 2 Tensor Cores per JFC chip, and 4 JFC chips per TPUv2 board . . . for a total 8 MXUs, or 131,072 LPHDR execution units." Ex. 1 ¶ 226.

"Under industry-standard metrics (*e.g.*, FLOP/s or FLOP/cycle), a person of ordinary skill in the art would understand that each MXU, together with its associated precision-reducer circuits, completes 16,384 floating-point operations—*i.e.*, LPHDR multiplications—per cycle, and thus contains 16,384 independent LPHDR execution units that execute in parallel." Ex. 1 ¶ 228.

[27] "Every clock cycle, MXUs of the accused TPUv2/3 devices complete 16,384 (128 x 128) LPHDR multiplication operations, with 16,384 separate output signals produced by 16,384 BF16 multiplier circuits." Ex. 1 ¶ 224.

"Thus, there 16,384 LPHDR execution units per MXU." Ex. 1 ¶ 225.

"Each TPUv3 device has 2 MXUs per Tensor Core, 2 Tensor Cores per DFC chip, and 4 DFC chips per TPUv3 board . . . for a total of 16 MXUs, or 262,144 LPHDR execution units." Ex. 1 ¶ 227.

"Under industry-standard metrics (*e.g.*, FLOP/s or FLOP/cycle), a person of ordinary skill in the art would understand that each MXU, together with its associated precision-reducer circuits, completes 16,384 floating-point operations—*i.e.*, LPHDR multiplications—per cycle, and thus contains 16,384 independent LPHDR execution units that execute in parallel." Ex. 1 ¶ 228.

[28] "In this report, I will frequently refer to a 'cycle' or a 'clock cycle' in explaining my opinions. Broadly speaking, computer circuitry operates at a pace that is dictated by a 'clock' circuit, which generates regular edges (or 'ticks') that allow different circuits to synchronize and coordinate their actions. The duration of a clock cycle is the period of time between one clock

35. Dr. Khatri opines that counting operations in lieu of physical structures is proper because "computer engineers quantify computational efficiency by counting the number of completed executions per unit time." Ex. 1 ¶ 233.[29]

Respectfully submitted,

Dated: April 28, 2023

By: */s/ Nathan R. Speed*

GREGORY F. CORBETT (BBO #646394)
gcorbett@wolfgreenfield.com
NATHAN R. SPEED (BBO #670249)
nspeed@wolfgreenfield.com
Elizabeth A. DiMarco (BBO #681921)
edimarco@wolfgreenfield.com
ANANT K. SARASWAT (BBO #676048)
asaraswat@wolfgreenfield.com
WOLF, GREENFIELD & SACKS, P.C.
600 Atlantic Avenue
Boston, MA 02210
Telephone: (617) 646-8000
Fax: (617) 646-8646

rising (or falling) edge and the next; it is – in effect – the time it takes to perform the smallest operation in a computer system. Some operations can be performed in a single clock cycle, while more complicated computing operations may require several clock cycles to complete. In modern computer architectures, it may take multiple clock cycles to perform an arithmetic operation." Ex. 1 ¶ 62 n.2.

[29] "Indeed, it is routine practice in all areas of engineering to maximally share resources (circuitry in this case) in order to reduce design cost. For this reason, computer engineers quantify computational efficiency by counting the number of *completed* executions per unit time." Ex. 1 ¶ 233 (emphasis in original)

ROBERT VAN NEST (*pro hac vice*)
rvannest@keker.com
MICHELLE YBARRA (*pro hac vice*)
mybarra@keker.com
ANDREW BRUNS (*pro hac vice*)
abruns@keker.com
VISHESH NARAYEN (*pro hac vice*)
vnarayen@keker.com
CHRISTOPHER S. SUN (*pro hac vice*)
csun@keker.com
ANNA PORTO (*pro hac vice*)
aporto@keker.com
DEEVA SHAH (*pro hac vice*)
dshah@keker.com
STEPHANIE J. GOLDBERG (*pro hac vice*)
sgoldberg@keker.com
KEKER, VAN NEST & PETERS LLP
633 Battery Street
San Francisco, CA 94111-1809
(415) 391-5400

MICHAEL S. KWUN (*pro hac vice*)
mkwun@kblfirm.com
ASIM BHANSALI (*pro hac vice*)
abhansali@kblfirm.com
KWUN BHANSALI LAZARUS LLP
555 Montgomery Street, Suite 750
San Francisco, CA 94111
(415) 630-2350

Matthias A. Kamber (*pro hac vice*)
matthiaskamber@paulhastings.com
PAUL HASTINGS, LLP
101 California Street
Forty-Eighth Floor
San Francisco, CA 94111
(415) 856-7000

Counsel for Defendant Google LLC

CERTIFICATE OF SERVICE

I certify that this document is being filed through the Court's electronic filing system, which serves counsel for other parties who are registered participants as identified on the Notice of Electronic Filing (NEF). Any counsel for other parties who are not registered participants are being served by first class mail on the date of electronic filing.

/s/ *Nathan R. Speed*
NATHAN R. SPEED